Контрольное задание

Компетенция «Машинное обучение и большие данные»

**Общее описание**

В эти три для мы с вами познакомимся с задачами машинного обучения, с которыми могут столкнуться специалисты в реальной жизни: от загрузки и предобработки данных и до вопросов комментирования кода.

**Модуль 1. Загрузка, предобработка и подготовка данных**

В данном модуле участники должны загрузить данные, выполнить предобработку, подготовить данные для последующего использования и подготовить отчет о проделанной работе.

Входные данные:

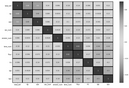
1. В архиве Stud\_IT\_input находятся следующие файлы: **matrix.png**, **Life\_expectancy\_dataset.csv**, **Aristotle.txt** и каталог **Energy\_consumption**.

Подробное описание задания:

1. Загрузите из папки **training\_datasets** файл **Life\_expectancy\_dataset.csv**.
   1. Для загруженного датасета рассчитайте среднюю продолжительность жизни для континента **Oceania.** Выведите полученное значение.
   2. Среди названий стран найдите страны, название которых содержит символы отличные от букв латинского алфавита, точек, запятых и пробелов. Для этих стран рассчитайте разницу между средней продолжительностью жизни женщин и мужчин. Выведите полученное значение.
   3. Среди названий стран найдите 3 записи с самыми длинными названиями по количеству символов. В данном случае все символы в строке учитываются как название страны. Среди данных стран найдите наиболее популярный континент. Название континента выведите.
2. Загрузите из папки **training\_datasets** файл **matrix.png**.
   1. Преобразуйте исходное изображение к виду, представленному на картинке ниже. Результат визуализируйте.  
      
   2. Преобразуйте исходное изображение к виду, представленному на картинке ниже. Для этого увеличьте контрастность изображения на фактор 2   
      Результат визуализируйте.



* 1. Преобразуйте исходное изображение к виду, представленному на картинке ниже. Для этого приведите изображение к серой цветовой гамме. **ВАЖНО! ЗАПРЕЩАЕТСЯ ИСПОЛЬЗОВАТЬ ВСТРОЕННЫЕ МЕТОДЫ ИЗ БИБЛИОТЕК ДЛЯ ПЕРЕВОДА В СЕРУЮ ГАММУ! РАЗРЕШАЕТСЯ ИСПОЛЬЗОВАТЬ МАТЕМАТИЧЕСКИЕ ПРЕОБРАЗОВАНИЯ НАД ДАННЫМИ**   
     Результат визуализируйте.



1. Загрузите из папки **training\_datasets** файл **Aristotle.txt.**
   1. Для загруженного датасета подсчитайте количество строк, содержащих слово «**человек»** в любом регистре. Выведите результат.
   2. Для всех строк в данных, подсчитайте, какие три слова (в любом регистре) находятся за словом человек чаще всего. Случай, когда слово человек находится в конце строки, учитывайте как наличие пустой строки '' после слова. Составьте список с этими тремя словами в порядке убывания их частоты встречаемости после слова человек Полученный список выведите.
   3. Далее вам предлагается построить простейшую таблицу [эмбеддингов](https://ru.wikipedia.org/wiki/%D0%92%D0%B5%D0%BA%D1%82%D0%BE%D1%80%D0%BD%D0%BE%D0%B5_%D0%BF%D1%80%D0%B5%D0%B4%D1%81%D1%82%D0%B0%D0%B2%D0%BB%D0%B5%D0%BD%D0%B8%D0%B5_%D1%81%D0%BB%D0%BE%D0%B2). Для этого вам сначала необходимо из каждой строчки удалить все символы, не являющиеся символами русского алфавита или пробелами, после чего разбить каждую строчку на отдельные слова. Далее вам необходимо подсчитать количество каждого из слов (в любом регистре) во всем корпусе текста и составить ранжированый список. И теперь для каждого слова в пару ему ставится целочисленное число, начиная с 1 для самого часто встречающегося слова.   
      Теперь организуйте этот процесс в виде функции, принимающей на вход любую строку текста, выполняющую предобработку и выдающую массив целочисленных значений (эмбеддингов), соответствующий данной строке на основании знаний о словах, полученных из **Aristotle.txt**. В случае, если в строке, поданной на вход функции встретилось слово, которого не было в **Aristotle.txt**, в таком случае предлагается сопоставить такому слову число 0.
   4. Передайте на вход функции строку «**Росатом - настоящая команда, где ценятся дружба, и каждый свободный работник волен решать, как он будет нести процветание в этот мир»**. Полученный на выходе эмбеддинг выведите на экран.
2. В папке **training\_datasets** откройте каталог **Energy\_consumption**. В нем вы найдете 13 файлов с потреблением электроэнергии от разных компаний в разное время. Загрузите файлы и объедините их вместе в один датафрейм.
   1. Постройте график , на котором будут нарисованы изменения во времени для каждого набора данных, на одном полотне с общими временной и мощностной осями.
   2. Постройте график для значений столбца **EKPC**. На основании графика сделайте выводы о характере сезонности и тренде параметра.
   3. Возьмите столбец с данными **EKPC**. Массив его значений разбейте на 5 фолдов для целей кросс-валидации Результирующие **5 пар train-test** данных. Визуализируйте пары на графиках, отрисовывая тренировочные данные синим цветом, а тестовые – красным.
3. Подготовить отчет по модулю (отчет должен быть в формате HTML/PDF на выбор, отчет должен быть структурирован, описывать ход работы, содержать комментарии).

Выходные данные по модулю:

1. Отчет в формате HTML.
2. Файл .ipynb, **полностью исполняемый и без выпадающих ошибок**.

Отчет должен иметь строго определенные название:

M1\_{lastname}\_{first name}

(например, M1\_Ivanov\_Ivan)